Evidence Reasoning for Event Inference in Smart Transport Video Surveillance

Xin Hong Wenjun Ma Yan Huang Paul Miller Weiru Liu Huiyu Zhou
Centre for Secure Information Technologies, School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, BT3 9DT, UK
{x.hong; w.ma; y.huang; p.miller; w.liu; h.zhou}@qub.ac.uk

ABSTRACT

In this paper we present a new event recognition framework, based on the Dempster-Shafer theory of evidence, which combines the evidence from multiple atomic events detected by low-level computer vision analytics. The proposed framework employs evidential network modelling of composite events. This approach can effectively handle the uncertainty of the detected events, whilst inferring high-level events that have semantic meaning with high degrees of belief. Our scheme has been comprehensively evaluated against various scenarios that simulate passenger behaviour on public transport platforms such as buses and trains. The average accuracy rate of our method is 81% in comparison to 76% by a standard rule-based method.

Keywords

Transport video surveillance, event detection, evidence reasoning

1. INTRODUCTION

The merging of advanced sensing technology and innovative pervasive computing has seen a rapid rise of interest in creating smart environments for our daily living. Smart surveillance, combining computer vision, networking and artificial intelligence (AI) technologies for extracting semantic information from distributed sensors [5] is a typical example. In particular, CCTV technology has been deployed to create secure transport corridors for the rapid transit of people [13].

Traditionally, CCTV systems operate in a passive mode, simply collecting enormous volumes of video data. Cameras are connected through a network to transmit video data to a central control room. A key to the success of active CCTV technology is the use of video analytics. Computer vision involves complicated techniques for interpreting the collected image data, however, by itself it is not sufficient to achieve the full capabilities required of intelligent surveillance. Further application-related high-level reasoning is needed. In this work, we develop an intelligent surveillance system for a transport application that is shown in Fig. 1. When we deploy video analytics, shown in the middle layer, visual features that a computer vision system can extract and handle without human intervention are mapped onto concepts (also called events) as perceived by humans [4], e.g. a male walking. The conceptual events are then used to further infer complex events that reveal sufficient semantics to human users for decision making.

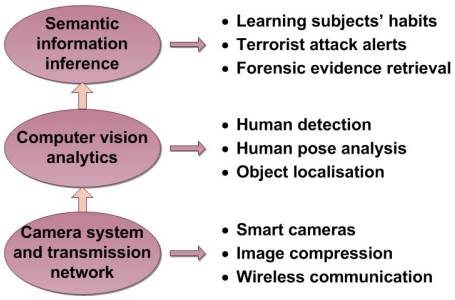


Figure 1: The three functional layers of an intelligent surveillance system

Simulating how humans process visual information is considered one of the most challenging tasks in AI. Complex events are usually made up of much simpler sub-events, which carry logical, spatial and temporal information. It is arguably very important to use structural information to describe events. In this work, we aim to recognise composite events literally occurring on buses based on observed passenger movements. We make use of the temporal and spatial relationship between human movements and/or event candidates in order to infer actual events. Dynamic environments with changing illumination onboard moving transport platforms, combined with the extremely large volumes of image data, may make event detection/recognition in surveillance video extremely challenging. Major problems include unreliable outputs of low-level computer vision analytics, such as incorrect object detection and tracking, varying renditions of identical events , the similar appearance of different events, and ambiguity in event definition [2][7]. Based on the Dempster-Shafer (DS) theory of evidence [3][17], we propose an evidential reasoning framework for event recognition in surveillance video. The proposed event network model can hierarchically represent structural relationships between composite events, atomic events, contexts and sensory evidence (i.e. outputs of low-level computer vision analytics). An embedded evidential reasoning system provides an ability to numerically represent uncertainty in relation to event recognition, inferring the occurrence of complex events with belief values, and making a decision on the most possible complex event that actually takes place. This paper extends our previous work [6], where we initially introduced the evidential event inference approach, in two ways: 1) we have refined the event network modelling and evidential event inference

algorithm, and 2) we present implementation details and more extensive experimental results.

The rest of this paper is organised as follows. In Section 2, we discuss related work, and highlight our contributions in comparison to it. Section 3 presents our evidential event network modelling and our proposed approach to recognising composite events with uncertainty, along with the main concepts of DS theory. Section 4 shows our system implementation in an indoors simulated bus environment, followed by the system evaluation. Finally, Section 5 concludes the paper with a discussion on future work.

2. RELATED WORKS

Over recent years, event recognition has received a lot of research attention, specifically in relation to video surveillance. There are roughly two different groups of approaches for event recognition developed up to date: deterministic and probabilistic [16]. Of the deterministic approaches, syntactic techniques such as context-free grammar [14], description methodologies such as Petri Nets [8], and logic-based approaches, such as Allen's temporal predicates [1], have been used to model events in real applications. These approaches can be used to describe the semantics of events, but lack an appropriate recognition step and do not consider uncertainty related issues [16]. For the probabilistic approaches [7][15], a probabilistic model is constructed from training data. In spite of the promising performance made within these methods, these probabilistic approaches are not capable of modelling complex events and have been restricted to very simple events/actions. In [11], the DS theory of permitting a framework to deal with imperfect information was used to extend a rule-based system for event inference. However, this event reasoning system has two major issues. One of them is that it was fairly difficult to obtain complete rules to cover all possible situations. The other one is that the ambiguity in the final results may be too high due to the way they assigned belief degrees given a shortage of sufficient rules.

In order to address the problems of a rule based system, while taking into account the two aforementioned issues, we propose here a hierarchical network to represent the structure of the event relationships, and apply DS theory to model uncertainty of the event inference. In our approach, we are able to bridge the semantic gap between the low level video data and high level human interpretation, and describe inference steps with better accuracy than a rule based system. Combining the power of DS theory and event reasoning network modelling, our approach is effective and can infer accurate and reliable events for various scenarios.

3. METHODOLOGY

3.1 Framework Architecture

Our proposed framework is composed of two main components at an upper and lower level, illustrated in Fig. 2. The two tiers of the recognition process integrates the advantages of both the computer vision techniques and the mechanisms of knowledge representation and reasoning. At the low level, human subjects are detected and video features are then extracted, using computer vision techniques, in order to provide low-level semantic components such as "a female face has been detected" and "a person has moved from the door towards the gang-way". The high level module of the system is designed to recognise significant events based on the semantic hierarchy of the events obtained from domain knowledge and human experts. At this level, the events of interest are recognised based on the information derived at the lower-level with degrees of belief.

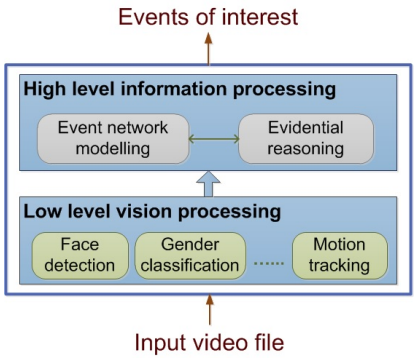


Figure 2: Evidential framework for event recognition

We have previously developed several video analytic modules for gender classification and movement tracking [12]. In this paper, we concentrate on investigating the high level event recognition of the proposed framework.

3.2 Event Network Modelling

Video surveillance has many application areas such as airport security and transport management. To meet these diverse goals, video surveillance systems must have extensive functionality. For example, for physical security, research work in video surveillance has been focused on the detection of anti-social behaviour in the last ten years. For transport management, the operator of a bus network as the end user of the video surveillance systems, needs to automatically record the passenger usage of their vehicles whilst obtaining dynamic information such as boarding events and seat selection.

At the second tier of the proposed framework, we derive semantic information of interest to the user from the video analytics outputs. The main purpose of video surveillance is to provide situational awareness of a specific place over a period of time. In the context of video surveillance, therefore, an *event* is an observation (or collection of observations) that has semantic perception. An event can be simple or complex, which is composed of simpler sub-events. To distinguish these two different concepts, we call the former an atomic event and the latter a composite event. An atomic event can be directly detected using sensors or video analytics. Atomic events can be aggregated to generate composite events which are more semantically meaningful.

To represent the hierarchical structure of the relationships between composite and atomic events, and the video analytic outputs, we propose an evidential network model for event inference.

DEFINITION 1. *An evidential event network (EEN) is a graph of upside-down tree* $EEN = (ND, EG, MM)$*, where:*

- ***ND*** = $\{n_1, ..., n_N\}$ *is a set of nodes representing events,*
- ***EG*** *is a set of edges over* ND*, each of which represents a close relation between the nodes at two consecutive layers,*
- ***MM*** *is a set of multi-valued mappings, which describe the compatibility relations between the node at the layer where an edge starts and the node at the layer where the edge ends.*

Fig. 3 shows the layout of an evidential event network (EEN). In an EEN, nodes can be categorised into three levels. The top level sits on a root node. At the bottom level, we have quite a few leaf nodes. Between these two levels, there are some nodes that can be divided into several sub-layers. Over the three levels, there exist

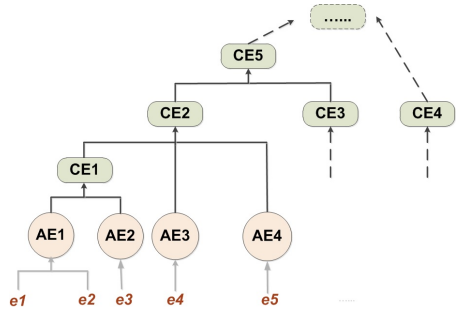


Figure 3: A general layout of evidential event networks

two types of nodes, associated with the level at which a node sits and the end of an edge to which the node is connected. A leaf node at the bottom level can be an atomic event, which is detected by a sensor, such as a seat pressure, or a video analytics module, such as face detection and tracking. A leaf node is always connected to the start of an edge. At the other end of the edge, we have nodes from the middle level. Middle level nodes are composite events derived from the connected atomic event nodes. Composite events at this level may be further connected together in order to form composite events at higher layers. On the topmost level of the EEN tree, there are composite events that are derived from atomic and/or composite events below, which are the events of interest to the end users. The hierarchical structure of an evidential event network reveals semantic relations between events, which are the foundation of evidential event inference developed below. This paradigm also helps in preventing redundancy by reusing the recognised atomic and composite events.

3.3 Event Representation with Uncertainty

Uncertainty is intrinsic to event recognition. Video sensors cannot provide complete information of an evolving scenario over time. In other words, video analysis modules have certain limitations with respect to providing correct visual information about a scene. During information processing, there is uncertainty in representing the relations between two events of interest. An event recognition system to be developed should be able to represent and infer useful information with uncertainty.

To manage uncertainty, we deploy DS theory [3][17], a generalisation of the traditional probability theory. DS theory describes the propositional space of possible situations for a given problem by a finite, non-empty set namely the *frame of discernment*, denoted as Θ. A unit of belief is distributed on Θ through a *mass function* m, satisfying (1) $m(\emptyset) = 0$ (2) $\Sigma_{A\subseteq\Theta} m(A) = 1$. From a mass function m, a belief function (Bel) and a plausibility function (Pls) can be derived, representing the degree of the justified and potential support given to A:

$$Bel(A) = \sum_{B\subseteq A} m(B) \text{ and } Pls(A) = \sum_{B\cap A\neq\emptyset} m(B).$$

For decision making, the pignistic transformation is proposed in [18]. If there exists a mass function $m(A), A \subseteq \Theta$, the pignistic probability $BetP$ for every element w of Θ can be calculated:

$$BetP(w) = \sum_{w\in A} \frac{m(A)}{|A|} \quad (1)$$

where $|A|$ is the number of elements of Θ in A. The pignistic probability is the counterpart of the subjective probability that quantifies the agent's beliefs according to a Bayesian probability theory.

To reflect the reliability of the source, a discount rate $r \in [0, 1]$ was

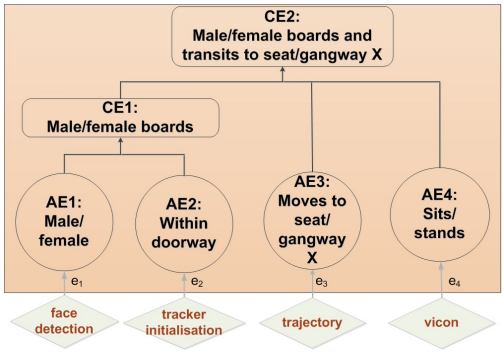


Figure 4: An Example of evidential event networks

introduced in [10]. The original mass function m from a source is discounted:

$$m^r(A) = \begin{cases} (1-r)m(A), & A \subset \Theta \\ r + (1-r)m(\Theta), & A = \Theta. \end{cases} \quad (2)$$

Here, we give a formal description of events in an evidential event network.

DEFINITION 2. *In an evidential event network, an event node is a tuple:*

n = (nType, Level, Date, Time, source, reliaR, vFrame, m)

where *nType* is the descriptor of an event node, such as *Female boards the bus*. *Level* informs whether the event is *Atomic* or *Composite*. *Date* and *Time* give the date and time an event occurs. *Source* denotes the unique identifier of a source, such as a seat sensor or a gender classification unit. Here, we use numerical numbers, e.g. 1, 2, to indicate a source. *reliaR* is the degree of reliability of a source. *vFrame* is the frame of discernment that holds all its values. *m* is a mass function on *vFrame*. It is worth mentioning that for a composite event, *source* and *reliaR* are not required. Below we use an example to demonstrate how to represent events and uncertain information are handled in an evidential event network.

SCENARIO 1. *A bus company is interested in monitoring the activities of passengers boarding a bus, e.g. how often female passengers board the bus and take a seat. A passenger is monitored after boarding a bus through the front door, transiting to a seat and sitting down. We define the composite event as "male/female boards, transits to seat/gangway X". Assume that we use camera A to capture the face of a passenger when (s)he boards the bus and to identify her/his gender, and camera B is used to detect the human body. Video analysis modules can support the detection of four atomic events: male/female, within/not-within the doorway, moves to seat/gangway X, sits/stands.*

Using domain knowledge, we know the relationships of the atomic and composite events for passengers. We can create an evidential event network as shown in Fig. 4. There are six event nodes created on this network: four atomic events - *AE1*, *AE2*, *AE3*, and *AE4*; two composite event - *CE1* and *CE2*. *AE1*, *AE2*, *AE3*, and *AE4* are detected by a video sensor. Valid values are assigned to all the tuple elements. For example, *AE1* is defined as n_{AE1} = $(AE1, atomic$, 28/02/2014, 14:20, 1, 90%, $\Theta, m)$ where $n_{AE1}.\Theta$ = $\{female, male, unknown\}$. *CE1* and *CE2* are inferred. For example, *CE1* is inferred from *AE1* and *AE2*, n_{CE1} = (*CE1*, *composite*, 28/02/2014, 14:20, ~, ~, Θ, m), $n_{CE1}.\Theta$ = {*(Female boards), (Male boards), (Nobody boards)*}.

Uncertainty buried in each component is defined as a mass distribution m. For an atomic event denoted as a leaf node of the event network, which is detected using low level computer vision modules, its mass value can be estimated based on the accuracy of the detection system. For a composite event, its mass distribution can be derived through an event inference process as detailed in the following section.

3.4 Evidential Event Inference

3.4.1 Evidential Reasoning

There are two main reasoning operators in DS theory. When two different frames hold compatibility relations, the mass function of one frame E can be *translated* to the other H via multivalued mapping Γ:

$$m(h_j) = \sum_{\Gamma(e_i)=h_j} m(e_i) \tag{3}$$

where $e_i \in E$, $h_j \subseteq H$ [9].

When two mass functions m_1 and m_2 are obtained from two independent sources over the same frame of discernment Θ, the consensus mass function m can be achieved by fusing them via Dempster's Rule of Combination,

$$m(C) = m_1 \oplus m_2 = (1-k)^{-1} \sum_{A \cap B = C} m_1(A) m_2(B) \tag{4}$$

where $k = \Sigma_{A \cap B = \emptyset} m_1(A) m_2(B) \neq 1$.The combination rule is both commutative and associative.

3.4.2 Event Inference Process

At the bottom level of an evidential event network, sensors or video analysis modules provide information for the atomic events as leaf nodes. Visual evidence can be used to infer the occurrence of the observed composite events. Evidence is propagated through the network to infer composite events at a higher level using evidential reasoning operations.

Composite event inference starts from obtaining event outputs from computer vision analysis modules and moves up within an evidential event network. The final output of the process is the composite event corresponding to the one that actually occurred. Algorithm 1 details the inference process.

Algorithm 1 Evidential event inference

Input: an event network EEN, mass functions on all atomic events obtained from visual analytics

Output: a composite event inferred

1: the mass functions of atomic events are discounted to take into account the reliability of visual analysis using Equation 2;
2: start from composite event nodes connected by all atomic event nodes below (so called child nodes);
3: **while** not reach the topmost node of the EEN **do**
4: translate mass functions of all child nodes into the node
5: using Equation 3;
6: combine the translated mass functions using Equation 4;
7: **end while**
8: calculate the $BetP$ of each element of the composite event node on the top using Equation 1;
9: select the element with the highest $BetP$ as the composite event inferred to output.

Now, we use scenario 1 to show how we calculate the mass distribution for a composite event of interest and make reliable recognition using the visual information.

SCENARIO 2. *(Scenario 1 continued) At 14:20, camera A captured a face of male, and camera B detected a passenger passing through the doorway. Within seconds, a human body was detected and tracked to the vicinity of seat 3. This is verified by the detection of a sitting posture. From the prior knowledge, we know that the gender classification unit has a 90% accuracy rate. The tracker has a reliability of 80%. The trajectories obtained from the tracker reveal that the passenger stops at a location close to seat 3 with a confidence of 60%, seat 4 with a confidence of 30% and any other seat with a confidence of 10%. Knowing the bus layout, we predict that the person is moving to a seat along the gangway because the end of his trajectory is near seat 3. Finally, the sit posture detection has a reliability of 92%.*

To explore what has taken place during this period of time, the inference procedure taking on the evidential event network in Fig. 4, starts from collecting evidence in the form of mass distributions on the nodes: $AE1$, $AE2$, $AE3$ and $AE4$, that are detected by the sensor and the video analysis components. For example, $n_{AE1}.m(\{male\}) = 0.9 * 1$. After that, $n_{AE1}.m$ and $n_{AE2}.m$ are translated onto n_{CE1} and combined to get $n_{CE1}.m$. Then, $n_{CE1}.m$, $n_{AE3}.m$ and $n_{AE4}.m$ are translated to n_{CE2} and combined to obtain $n_{CE2}.m$. On $CE2$, $BetP$ is calculated, and as a result, the composite event "MALE BOARDS THE BUS AND TRANSITS TO SEAT 3"is generated.

Our evidential event network approach has two advantages in comparison to the method presented in [11]:

(1) *Robust scalability*. Consider the example of passengers boarding bus and taking a seat shown in Scenario 1 and 2. Only one evidential event network shown in Fig. 4 is needed here to describe the relations of the atomic and composite events in order to infer a composite event "A PERSON OF MALE/FEMALE GENDER BOARDS AND TRANSITS TO A SEAT". However, in [11], an inference rule must be specified for each event of interest. With this example, a set of rules is required such as "FEMALE BOARDS $\wedge$ TRANSITS TO SEAT 1 $\rightarrow$ FEMALE BOARDS AND TRANSITS TO SEAT 1". Suppose there are 32 seats on the bus, with two genders their method requires 64 rules to derive the necessary results. This leads to two problems: (a) Reduced efficiency when the system handles complex situations with a large number of composite events. (b) Such a rule based system requires more experts with extensive domain knowledge to construct a workable system.

(2) *Less ambiguity degree in inferring the event of interest*. With the example, assume that we have $m(\{$*male broads*$\}) = 0.9$, $m(\{$*male broads, female broads*$\}= 0.1$; $m(\{seat3\})= 0.6$, $m(\{seat4\})= 0.3$, $m(\{all\ seats\})= 0.1$. Using our proposed method, we can obtain the mass functions: $m(\{$*male broads and transits to seat* $3\})= 0.54$ (in short $m(mbts3) = 0.54$), $m(\{$*male broads and transits to seat* $4\ \})= 0.27$ (in short $m(mbts4)= 0.27$), $m(\{$*male broads and transits to any seat*$\})= 0.09$ (in short $m(mbts)= 0.09$), $m(mbts3, fbts3)= 0.06$, $m(mbts4, fbts4)= 0.03$, $m(\Theta)= 0.01$. Thus, with a 10-seats bus, our approach leads to $BetP(mbts3)= 0.5805$. Now, using the approach reported in [11], two rules will be involved: "MALE BOARDS $\wedge$ TRANSITS TO SEAT 3 $\rightarrow$ MALE BOARDS AND TRANSITS TO SEAT 3"; "MALE BOARDS $\wedge$ AND TRANSITS TO SEAT 4 $\rightarrow$ MALE BOARDS AND TRANSITS TO SEAT 4". Thus, [11] results in $m(mbts3) = 0.54$, $m(mbts4)= 0.27$, $m(\Theta)= 0.19$. With a bus of the same size, their approach achieves $BetP(mbts3)= 0.5495$. Our method achieves a higher confidence than [11] in the final output.

Table 1: Event list

Alias	Title
MBTSt	Male boards and transits to seat X
FBTSt	Female boards and transits to seat X
PEX	Person exits the bus
PCS	Person changes seat

4. EXPERIMENTS

We have evaluated the performance of our proposed event recognition framework using a dataset collected from a simulated bus scenario. The goal of video surveillance is to effectively identify human activities on buses and recognise criminal and anti-social behaviours. We start from tracking passengers who board the bus and continuously track them as they move, sit and later alight from the bus. Within this context, there are four broad human activities: boarding, moving, sitting and alighting. In the experiments, we are particularly interested in evaluating the system performance with respect to the detection of the four composite events as listed in Table 1.

4.1 Set-up

In our simulated bus environment, we use a Panasonic camera WV-NP244 (camera A) to monitor the entry area, and an AXIS M31-R camera (camera B) to monitor the saloon. The room door-way area simulates the bus door. Camera A is positioned so that it can capture a passenger's face as (s)he enters the bus. The imagery from camera A is provided as the input to a face detection module with a gender classification tool. In this simulation, there are 17 seats for passengers to sit and a gangway (the corridor between two seat columns) for passengers to stand. Camera B covers the whole range of seats, the gangway and the doorway. The imagery from camera B is provided as the input to a human detection and tracking module [12]. A Vicon sensor is also worn by each passenger to provide ground truth motion. The Vicon motion capture system can provide millimetre-accurate position information of each passenger. Fig. 5 shows the schematic of the simulated bus area, including the position of the cameras. The region R is designed to be one meter away from the entrance/exit of the bus and is used to determine passengers boarding and alighting.

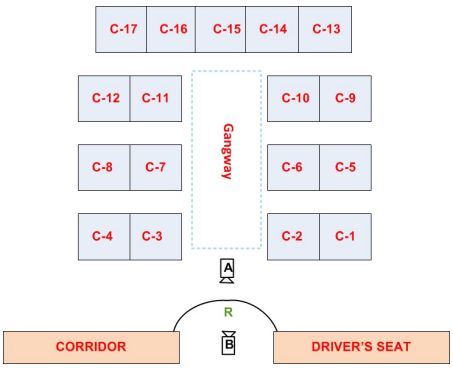


Figure 5: A floor plan schematic of the simulated bus area

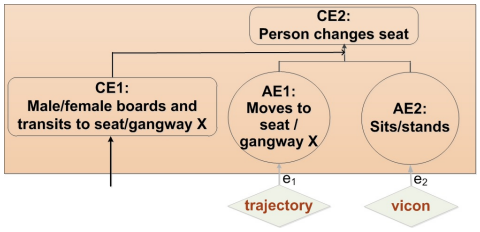
4.2 Methodology

The frame rate of the captured video is set to be 30 FPS. Video sequences are fed into our in-house face detection module and the tracking system to produce atomic events. Detected atomic events are then used to infer the composite events of interest.

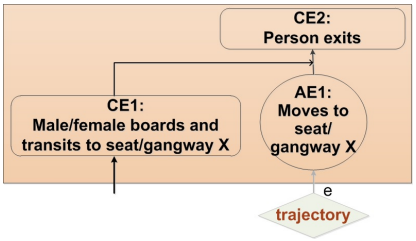
Table 2: Statistical results

Event	MBTST	FBTST	PEX	PCS	accuracy
Groundtruth	9	9	18	6	100%
Our method	4	7	17	6	81%
Rule-based	2	7	17	6	76%

Knowing the topology of the bus environment and the activities of interest, we constructed three evidential event networks as shown in Figure 4 and 6. The visual evidence is used to infer composite events on the event networks. Fig. 7a - 7c show an inference



(a) Person changes seat



(b) Person exits the bus

Figure 6: Evidential event networks

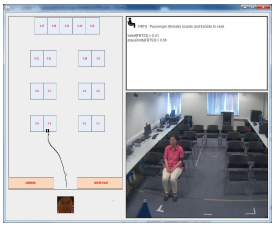
process used in the experiments. These are screen shots of the graphical interface of our event inference system at three different instances: instance 1 - a female boards the bus and transited to seat 3; instance 2 - the person changed to seat 5; instance 3 - the person exited the bus. The left side is the plan view of the bus area with trajectories obtained by our in-house tracking system [12]. The bottom area of the right-hand side shows one of the captured views. The top right-hand side area shows the events detected in the video, along with the belief and plausibility of an event that was automatically recognised by the proposed system.

4.3 Results

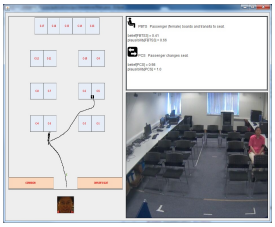
Six passengers (three female and three male adults) participated in the experiments. Video recordings include eighteen sequences, each of which lasts around thirty seconds, of a single passenger walking around in the bus. Three scenarios were recorded. In the first group, a passenger enters the bus, selects a seat and then exits the bus. In the second group, a passenger enters the bus, takes a seat, changes to another seat, and then exits the bus. In the final group, a passenger enters the bus, walks towards the back row, turns around, sits down, and finally exits the bus.

Table 2 shows the amount of correctly inferred instances for each composite event and the average accuracy rate by our method and a rule-based method as a bench-mark. In addition, we show the ground truth provided by the Vicon system.

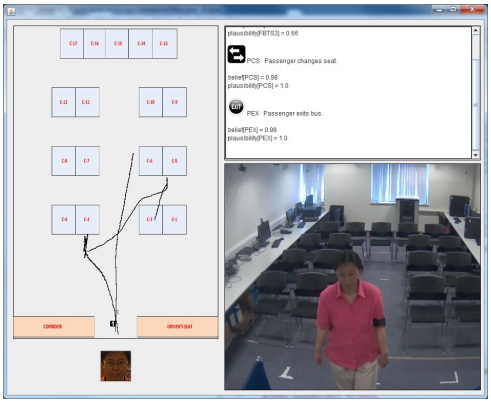
We used the rule-based system as the bench mark method to compare with our method. The results show that both methods per-



(a) Instance 1



(b) Instance 2



(c) Instance 3

Figure 7: Passenger's event reasoning demonstration

formed equally well on the inference of FBTSt, PEX and PCS events. However, our method was able to produce more accurate results than the rule-based system on inferring the MBTSt events, where many of the event instances have suffered from noisy video analysis with higher uncertainty, in particular, gender classification. This indicates the power of our method in event reasoning.

5. SUMMARY AND FUTURE WORK

In this paper we have presented a framework of representing the structural knowledge of events and reasoning about complex events based on the outputs of low-level video analytics. The proposed approach takes into account the uncertainty in the different stages of event representation, recognition and low-level video analytics. The proposed framework has provided reliable recognition results of complex scenarios using numerical belief measures.

The experiments show that the proposed framework is able to recognise complex events, not only when the tracking results were perfect, but also when the tracking process contained errors. Further experimental evaluations against more challenging scenarios in comparison with other state of art methods will be carried out in future work. We are also developing a mechanism to accommodate the temporal relations of evidential event network representation and intend to investigate the feasibility of deploying more video devices, or other sensing devices, to improve event recognition performance.

6. ACKNOWLEDGEMENTS

This work is partially supported by the CSIT project funded by UK EPSRC under the grant EP/H049606/1, Invest NI and various industrial partners.

References

[1] J. Allen and G. Ferguson. Actions and events in interval temporal logic. *J. Logic Comput.*, 4(5):531–579, 1994.

[2] F. Bashir, A. Khokhar, and D. Schonfeld. Real-time motion trajectory-based indexing and retrieval of video sequences. *IEEE Trans. Multimedia*, 9:58–65, Jan 2007.

[3] A. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Ann. Stat.*, 28:325–339, 1967.

[4] T. Geerinck, V. Enescu, I. Ravyse, and H. Sahli. Rule-based video interpretation framework: application to automated surveillance. In *Procs. of ICIG*, pages 341–348, 2009.

[5] A. Hampapur, L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, S. Pankanti, S. A., C. Shu, and Y. Tian. Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking. *IEEE Sig. Proc. Magazine*, 22(2):38–51, 2005.

[6] X. Hong, Y. Huang, W. Ma, P. Miller, W. Liu, and H. Zhou. Video event recognition by Dempster-Shafer theory. In *Procs. of ECAI*, 2014.

[7] G. Lavee, E. Rivlin, and M. Rudzsky. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. *IEEE T. Syst. Man. Cy. C*, 39(5):489–504, 2009.

[8] G. Lavee, M. Rudzsky, and E. Rivlin. Video event modelling and recognition in generalized stochastic petri nets. *IEEE T. Circ. Syst. Vid.*, 20(1):102–118, 2010.

[9] W. Liu, J. Hughes, and M. McTear. Representing heuristic knowledge in the DS theory. In *Procs. of UAI*, pages 182–190, 1992.

[10] J. Lowrance, T. Garvey, and T. Strat. A framework for evidential-reasoning systems. In *Procs. of AAAI*, pages 896–903, 1986.

[11] J. Ma, W. Liu, and P. Miller. Event modelling and reasoning with uncertain information for distributed sensor networks. In *Procs. of SUM*, pages 236–249. IEEE Press, 2010.

[12] N. McLaughlin, J. Martinez-del Rincon, and P. Miller. Online multiperson tracking with occlusion reasoning and unsupervised track motion model. In *Procs. of AVSS*, pages 37–42, 2013.

[13] P. Miller, W. Liu, F. Fowler, H. Zhou, J. Shen, J. Ma, J. Zhang, W. Yan, K. McLaughlin, and S. Sezer. Intelligent sensor information system for public transport: To safely go ... In *Procs. of AVSS*, pages 1–12, 2010.

[14] D. Moore and I. Essa. Recognizing multitasked activities from video using stochastic context-free grammar. In *Procs. of AAAI*, pages 770–776, 2002.

[15] R. Romdhane, B. Boulay, F. Bremond, and M. Thonnat. Probabilistic recognition of complex event. In *Procs. of ICCVS*, pages 122–131, 2011.

[16] J. SanMiguel and J. Martinez. A semantic-based probabilistic approach for real-time video event recognition. *Computer Vision and Image Understanding*, 116:937–952, 2012.

[17] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[18] P. Smets. Constructing the pignistic probability function in a context of uncertainty. In *Procs of UAI*, pages 29–40, 1990.